5-10-2014

High Order Statistics of Natural and Manmade Sounds

Rahul Narayan
University of Connecticut - Storrs, rahul.narayan@uconn.edu

High Order Statistics of Natural and Manmade Sounds

Rahul Narayan

B.Tech., Institute of Engineering & Technology, Agra, 2008

A Thesis

Submitted in Partial Fulfillment of the

Requirements of the Degree of

Master of Science

at the

University of Connecticut

2014

APPROVAL PAGE

Master of Science Thesis

High Order Statistics of Natural and Manmade Sounds

Presented by

Rahul Narayan, BTech.

Major Advisor _____

Dr. Monty Escabi

Associate Advisor _____

Dr. Heather Read

Associate Advisor _____

Dr. Krystyna Gielo-Perczak

University of Connecticut

2014

DEDICATION

This work is dedicated to my family, my advisor Dr. Monty Escabi, my associate advisors, faculty and staff at university of Connecticut. I would like to thank each one of the above for their patience, support and encouragement.

ACKNOWLEDGEMENTS

I would like to thank my Advisor Dr. Monty Escabi for his guidance, support and encouragement throughout the course of this research. I would also like to thank my committee members, Dr. Heather Read and Dr. Krystyna Gielo-Perczak for their support.

Finally I would like to thank my family, friends, staff and faculty at UConn for their constant support and understanding.

Table of Contents

1 Introduction 1

2 Methods 5

2.1 Categorical Sound Ensemble 5

2.2 Auditory Filterbank Model 8

2.3 Time varying Amplitude distribution 10

2.4 Joint Intensity and Contrast Statistics 12

2.5 Using Statistics for Sound Discrimination 12

3 Results 15

3.1 Time varying Statistics – Contrast and Intensity 15

3.2 Contrast and Intensity Distributions 21

3.3 Discrimination Performance Based on Observation Time 23

3.4 The Role of Contrast and Intensity for Sound Category Discrimination Performance 25

4 Discussion 32

4.1 Discussion 32

5 References 35

ABSTRACT

High Order Statistics of Natural and Manmade Sounds

Rahul Narayan

Major Advisor: Dr. Monty Escabi

Environmental sounds, both man-made and natural, vary on multiple time and frequency scales generating a large range of temporal, spectral and amplitude modulations that are evident in the high-order statistics of the sound spectrogram. Healthy hearing humans perceive high-order statistical regularities and use this information to categorize and discriminate sounds. This paper tests the hypothesis that biologically motivated sound statistics can enable/enhance discrimination and identification of sound categories from a computational standpoint. A large catalogue of natural and man-made sounds and their associated high-order contrast and intensity statistics were developed, and the information carrying content of each statistic for sound recognition and discrimination was measured. Bayesian classification and signal detection theory were applied to the sound database to identify statistics that can be used to categorize sounds and to test discrimination limits amongst sounds or categories. The catalogue will be deployed as an online database available to researchers and scientists.

LIST OF FIGURES

FIGURE 1: DENDROGRAM OF TEXTURES 6

FIGURE 2: DENDROGRAM OF ANIMAL VOCALIZATIONS 6

FIGURE 3: EXAMPLE OF DATABASE ENTRIES 7

FIGURE 4: DECOMPOSITION OF SOUNDS 8

FIGURE 5: SPECTRO-TEMPORAL ENVELOPE OF HUMAN SPEECH. 16

FIGURE 6: TIME VARYING AMPLITUDE DISTRIBUTIONS FOR ANIMAL VOCALIZATIONS 17

FIGURE 7: SPECTRO-TEMPORAL ENVELOPE OF A WATER SOUND. 18

FIGURE 7: TIME VARYING AMPLITUDE DISTRIBUTIONS FOR BACKGROUND SOUNDS 19

FIGURE 8: PARAMETERIZING AMPLITUDE DISTRIBUTION OF HUMAN SPEECH 20

FIGURE 9: PARAMETERIZING AMPLITUDE DISTRIBUTION OF PARROTS, WATER 20

FIGURE 10: INTENSITY VS CONTRAST STATISTICS FOR HUMANS 22

FIGURE 11: INTENSITY VS CONTRAST STATISTICS FOR MUSIC (CLASSICAL) 22

FIGURE 12: INTENSITY VS CONTRAST STATISTICS 23

FIGURE 13: CLASSIFIER PERFORMANCE PARROTS VS CATS 24

FIGURE 14: CLASSIFIER PERFORMANCE. PARROTS VS HUMANS 25

FIGURE 15: CONFUSION MATRIX FOR 15 SOUND CATEGORIES 26

FIGURE 16: GRAPHICAL REPRESENTATION OF CONFUSION 27

FIGURE 17: CLASSIFIER PERFORMANCE INCREASES WITH NUMBER OF POINTS 28

FIGURE 18: CLASSIFIER PERFORMANCES AT 256 SAMPLE POINTS 29

FIGURE 19: CLASSIFIER PERFORMANCE FOR MEAN ALONE AT 256 SAMPLE POINTS 30

FIGURE 20: CLASSIFIER PERFORMANCE FOR SD ALONE AT 256 SAMPLE POINTS 30

FIGURE 21: SUBSTANTIAL INCREASE IN CLASSIFIER PERFORMANCE AFTER COMBINING STATISTICS 31

CHAPTER 1

INTRODUCTION

Environmental sounds, both man-made and natural, vary on multiple time and frequency scales generating a large range of temporal, spectral and amplitude modulations that are evident in the high-order statistics of the sound spectrogram[1-4]. Healthy hearing humans perceive high-order statistical regularities and use this information to categorize and discriminate sounds[4]. We hypothesize that biologically motivated sound statistics can enable/enhance discrimination/identification of sound categories from a computational standpoint.

Sound textures are distinguished by temporal homogeneity, suggesting they could be recognized with time-averaged statistics[4]. They processed real-world textures with an auditory model containing filters tuned for sound frequencies and their modulations, and measured statistics of the resulting decomposition. The results showed that real-world textures can be synthesized from random noise by matching statistics of the decomposition. The realism and recognizability of novel sounds synthesized to have matching statistics were tested by playing these sounds to humans and asking them to rate the sounds based on reality. They found out that simple statistics such as the power spectrum failed to produce compelling synthetic textures but including high-order statistics such as correlations between channels produced identifiable and natural-sounding textures.

We set out to find if contrast and intensity statistics can be used to discriminate and classify sound categories. Analogous to visual contrast, sound contrast is defined by the relative range of sound level fluctuations, such that sounds that span an extensive range of sound levels have high contrast. Rather than using human subjects to classify sounds[1], computational analysis using a Bayes classifier was used to test the role contrast and intensity statistics play in discrimination phenomena. Prior studies demonstrate that central auditory neurons can respond selectively to high-order sound statistics including the sound contrast and statistics related to variations in the sound pressure level [1, 5-10]. Presently there is no comprehensive theory for how the brain encodes and represents high-order statistical regularities in sound, and in particular the role statistics play in sound recognition phenomena. In this study the high-order statistics of contrast and intensity were used. The contrast of natural sounds is described by the probability distribution of relative amplitudes (i.e., in units of dB) because neurons are highly sensitive to proportional fluctuations, not just simply the extreme values[1].

There are a variety of applications for sound detection and discrimination based high-order sound statistics. They can be broadly classified into (1) Technical, (2) Scientific, and (3) Clinical applications.

Technical Applications

Current Speech Recognition techniques are based on feature identification. Modern speech and sound recognition systems do not account for statistical regularities in complex sound mixtures. Behavioral studies have found that statistical regularities

contribute substantially to perception and discrimination performance in human listeners, and there is need to understand how the brain deals with such sound properties. There can be military applications for an efficient sound recognition system based on sound statistic and sound discrimination by separating out noise and other unwanted sounds.

Clinical Applications

Clinical treatments, including prosthetics and aids for auditory processing deficits, are not designed to dynamically or otherwise manipulate statistical regularities of sounds. Statistical regularities contribute substantially to perception and discrimination performance. Many environmental sounds such as those from a busy street or a crowded room are aptly described by high-order statistics[4]. Since such sounds often interfere during speech and sound recognitions tasks they can present significant challenges for the hearing impaired and for speech recognition technologies. Hearing aids, cochlear implants that take into account sound statistics from sound decomposition will be able to tackle these challenges. Approximately 17 percent (36 million) of American adults report some degree of hearing loss - National Institute on Deafness and Other Communication Disorders (NIDCD) Only 1 out of 5 people who could benefit from a hearing aid actually wears one.

Scientific Applications

The Sound discrimination techniques based on high-order statistics can be used for scientific studies and research. For example ARBIMON -Automated Remote Biodiversity Monitoring Network is a web based network for storing, sharing, and

analyzing acoustic information recorded from different environments including rainforests, urban settings etc.[11]. It uses this acoustic information to understand current patterns of land change. Such studies would benefit from the system.

CHAPTER 2

METHODS

a. Categorical sound ensemble

The objective is to develop a catalogue of high-order statistics from large ensembles of natural and man-made sounds and to quantify the information carrying content of each statistic for sound recognition and discrimination. We hypothesize that biologically motivated sound statistics can enable/enhance discrimination/identification of sound categories from a computational standpoint. It is particularly important to include manmade sounds, including music, in the catalogue because man-made background sounds present substantial challenges for speech recognition systems and the hearing impaired. Also, knowing the statistics of music could be beneficial for coding and compression. It is essential to include animal vocalizations and speech because knowledge of their statistics could significantly benefit speech recognition and auditory prosthetic technologies. Natural sounds are obtained through the Cornell Macaulay Library of Ornithology (http://macaulaylibrary.org/), other commercial sources. Man-made sounds (e.g., machines, music etc) will be obtained from commercially available media. The catalogue will be deployed as an online archive available to researchers and scientists.

Development of a hierarchical sound catalogue: Sounds are classified into hierarchical categories including vocalizations, background/environment sounds, and music. These are subcategorized according to species for vocalizations (e.g., human, non-

human primate, amphibians, birds etc.) or categories such as the acoustic source for background sounds (e.g., water, wind, speech babble, etc.) or man-made sounds (e.g., motorized sounds, impulsive sounds such as a hammer, etc.). An example dendrogram (i.e., cluster tree) representation of such categories is shown in Figures 2.0, 2.1.

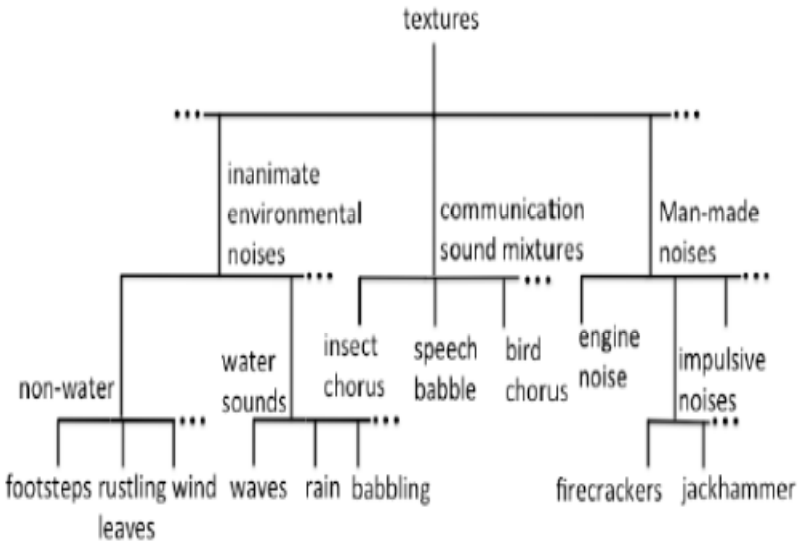

Figure 2.0: Dendrogram of Textures

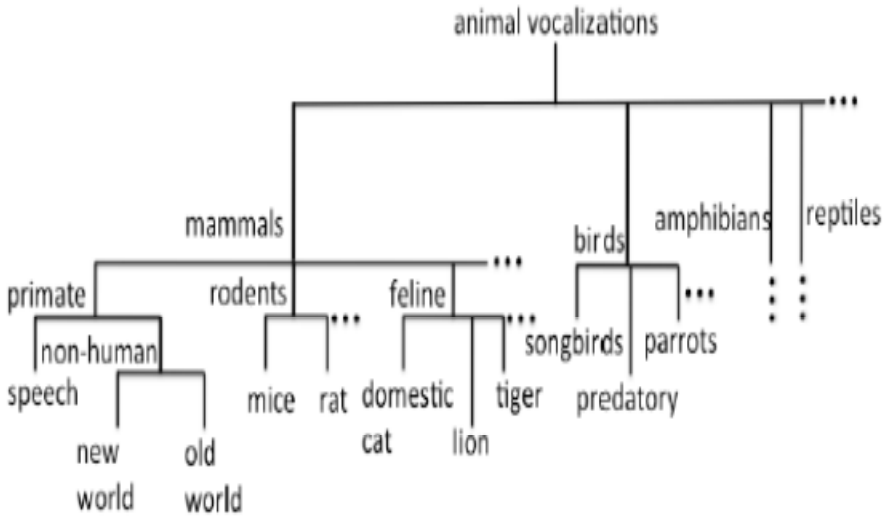

Figure 2.1: Dendrogram of Animal Vocalizations

17 CDs were analyzed which contained 1000+ individual sound tracks. More than 600 species of animals/birds were included in the study. They included human speech, animal vocalizations, background sounds and manmade sounds. Each track was listened to and clean segments and soundscape segments for each track was noted in an excel database. The species names, duration of the track, species category were also noted. A sample database is shown below:

Track	Type		Species														Background sounds				Comments :)
	Vocal	Texture	Birds				Amphibians			Mammals				Fish	Insects					Mechanical	
			Song	Predatory	Parrots	Others	Frogs	Toad	Others	Humans	Primates		Others				Birds	Insects	Others	Movemnt	Machine
											Old Wld.	New Wld.									
1	Y	Y				y											y	y		y	Satin Bowerbird
2	Y	Y							y								y	y		y	American Alligator
3	Y	Y											y						y		Jaguar
4	Y	Y				y													y		Super Lyrebird
5	Y	Y				y													y		Three-humped Treehopper
6	Y					y													y		Yellow Buffalo Treehopper
7	Y	y				y											y			y	Capercaillie
8	Y	Y											y					y			Franquet's Epauletted Bat
9	Y	y											y						y		Bowhead Whale
10	Y	y											y						y	y	Humpback Whale
11	y	y					y											y			Knudsen's Frog
12	Y	Y					y											y			Tungura Frog
13	Y	Y					y											y			Smoky Jungle Frog
14	Y	Y				y											y		y		Club-winged Manakin
15	Y	Y				y												y			Black-and-gold Cotinga

Figure 2.3: Example of database entries

b. Auditory filterbank model

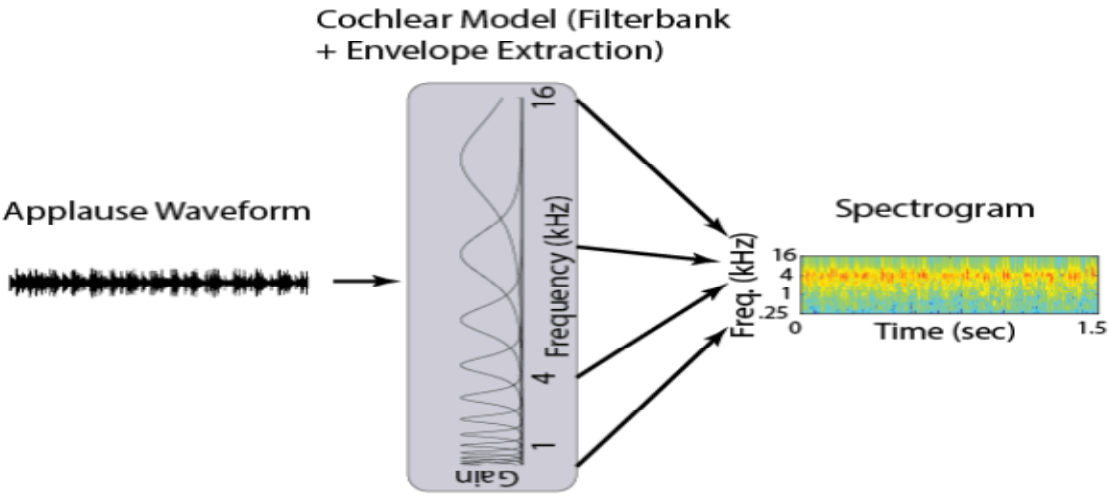

Fig. 2.4: Physiologically plausible decomposition of sounds into spectrotemporal acoustic elements. An example Applause sound waveform (A) is passed through a cochlear filter model (B) resulting in a spectrogram representation (C, time vs. frequency; red=high, blue=low power).

An auditory filter bank is used to divide the spectrum into components in a fashion similar to the way the hair cells in the cochlea respond to auditory stimuli. Engineers can use this information to determine which sounds are masked and which ones are audible. To minimize errors, the filter bank should be as accurate as possible.

Within the cochlea, sound waves travel through a fluid and excite small hair cells along the basilar membrane. High frequency tones excite hair cells near the oval window whereas low frequency tones affect hair cells near apical aspect of the cochlea. However, a single-frequency tone does not merely enervate a small number of hair cells. A simple tone excites hair cells most at a particular point, but it also excites surrounding hair cells (to a lesser extent).

Natural sounds and white noise were decomposed into their spectral and temporal components with a physiologically motivated filterbank that resembles the filtering characteristics of the peripheral auditory filters in mammals and perceptual filtering characteristics of humans. The filterbank model is similar to that described by Rodriguez et al [2]. Sounds were initially decomposed by a bank of tonotopically arranged filters into a spectrotemporal representation that mimics the spectral decomposition performed by the cochlea. Filter center frequencies were arranged according to the frequency position function of the cochlea over a range covering 250 Hz to 14 kHz, and filter bandwidths were selected according to the perceptual critical bandwidths. Sounds waveforms were decomposed according to:

$$s_k(t) = h_k(t) * s(t)$$

where h_k (t) is the impulse response of the k-th filter channel centered about the frequency f_k, * is shorthand for the convolution operator, and s(t) is the sound waveform.

All sounds were first filtered with an array of third-order (n = 3) gammatone filters (Irino and Patterson, 1996) with impulse response functions of the form h_k (t) = t_{n-1} · cos (2πfkt) · e(−2*π*b (fk)*t) where f_k represents the frequency of the kth filter and b(f_k) the filter bandwidth. The spectrotemporal envelope (s(t,x_k)) of each sound was obtained by passing the sound through the auditory filterbank and subsequently computing the magnitude of the analytic signal for each frequency channel:

$$s(t,x_k)= s_k (t)=| h_k (t)*s(t)+i\cdot H\{ h_k (t)*s(t)\}|.$$

Here s(t) is the input sound, s_k (t) is the extracted envelope for the kth channel, * represents the convolution operator, x_k is the frequency variable in octaves, and H{·} is the Hilbert transform. Filter center frequencies (f_k) were logarithmically spaced (1/8 octave spacing) between 200 Hz and 16 kHz and filter bandwidths [$b(f_k)$] were chosen to follow perceptual critical bandwidths: $b(f_k) = 25 + 75 \cdot [1 + 1.4 \cdot f_k^2]^{0.69}$. The temporal modulations within each frequency channel were then band limited to 800 Hz by filtering the temporal envelope with a b-spline lowpass filter. This upper limit was chosen because neurons in the central auditory system (e.g., inferior colliculus) do not phase-lock beyond this range.

c. Time varying amplitude distribution

Contrast, the range of amplitude excursions in the spectrogram of sounds, can enhance perceptual discrimination and identification. Such perceptual advantages may arise from neural sensitivities to contrast that are found in IC and AC[1]. Vocalizations and background sounds can be categorized as having high and low contrast, respectively, which could aid in sound detection. The contrast of natural sounds is best described by the probability distribution of relative amplitudes (i.e., in units of dB) because neurons are highly sensitive to proportional fluctuations, not just simply the extreme values[1]. Furthermore, intensity discrimination and modulation detection correlate best with proportional changes in amplitude.

Visual contrast is defined as the percent deviation relative to the mean intensity of a spatial sinusoid grating. Mathematically it is expressed as C=(Imax−Imin) (Imax+Imin) where Imax and Imin correspond to the maximum and minimum stimulus intensities [12, 13]. In the auditory literature the analogous quantity is the modulation depth or modulation index,b=(Imax−Imin) / Imax. Such a description suffices for the case of sinusoidal, square wave, and other simple stimulus gradations since these waveforms are fully specified by their minimum and maximum intensities. For natural signals, where the amplitude gradations can cover several orders of magnitude, such descriptions fail to fully characterize amplitude fluctuations since they only take into account the minimum and maximum envelope intensities. They do not tell us anything about intermediate values and higher−order amplitude statistics of the modulation signal. To overcome this we adopt a more general definition of contrast to denote the probability distribution of the relative amplitude gradations.

Many sounds also vary dynamically over time and for this reason the time-varying amplitude distribution was measured for each sound. The distribution is defined as $p_k(s)$, where s is the sound level in dB and t_k is the time of the k-th measurement. For each sound, the distribution is measured discretely using non-overlapping time-intervals of 50 msec.

d. Joint intensity and contrast statistics

To quantify the observed contrast dynamics for the various sound ensembles, the time−dependent amplitude distribution was parameterized by computing its time−dependent mean value, and its standard deviation. For all sounds in a given ensemble the joint histogram for these quantities was computed. The joint histogram was normalized so that its cumulative sum gives unity probability

e. Using statistics for sound discrimination

Bayesian classification and signal detection was used for sound discrimination. A naïve Bayes classifier was used which makes the assumption of independence between features. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. Abstractly, the probability model for a classifier is a conditional model

$$p\left(C|F_1,\cdots,F_n\right)$$

over a dependent class variable C with a small number of outcomes or *classes*, conditional on several feature variables F_1 through F_n. The goal is to maximize the likelihood of a particular class given the observation of the feature variables. The problem is that if the number of features *n* is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, this can be written

$$p\left(C|F_1,\cdots,F_n\right)=\frac{p\left(C|F_1,\cdots,F_n\right)p\left(F_1,\cdots,F_n|C\right)}{p\left(F_1,\cdots,F_n\right)}$$

In plain English, using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior}=\frac{\text{prior}\times\text{likelihood}}{\text{evidence}}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on *C* and the values of the features F_i are given, so that the denominator is effectively constant. Now the "naive" conditional

independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$ given the category *C*. Thus under the independence assumptions, the conditional distribution over the class variable *C* is:

$$p(C|F_1,\cdots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$$

where the evidence $Z = p(F_1,\cdots,F_n)$ is a constant scaling factor if the feature variables are known. Similiarly, for a given experiment where the class probability is known a priori, $p(C)$,is constant. Thus, for the Naïve Bayesian classifier, the selected class is the one that maximize the likelihood function $\prod_{i=1}^{n}p(F_i|C)$.

Signal Detection Theory

Detection theory, or signal detection theory, is a means to quantify the ability to discern between information-bearing patterns (called stimulus in humans, signal in machines) and random patterns that distract from the information (called noise, consisting of background stimuli and random activity of the detection machine and of the nervous system of the operator). In the field of electronics, the separation of such patterns from a disguising background is referred to as signal recovery.

CHAPTER 3

RESULTS

Time-varying statistics of the sound contrast and sound pressure level where measured for multiple sound categories at intervals of 50 msec[1]. These relative short intervals are chosen because perceptual integration of intensity and contrast occurs within a time scale of ~50-200 msec[1]. After generating sound catalogue with the associated statistics, we tested the hypothesis that these statistical features can be used to identify and/or discriminate sound categories. Bayesian classification was applied to the sound ensembles and the sound discrimination performance for contrast and sound level statistics was evaluated.

a. Time varying contrast and intensity statistics of natural sound ensembles

For each sound in the catalogue, we measured the time-varying distribution of spectrogram amplitudes at intervals of 50 msec. An example is shown for speech where we measured the amplitude distribution at three distinct time points from the auditory spectrogram (Fig. 3.1, at 0.7, 0.9 and 1.4 sec). The probability distributions of amplitudes (in dB) are shown at the selected time instants by measuring the amplitudes over a 50 msec window. As can be seen, the spectrogram amplitudes are highest about 0.7 sec and intermediate at 0.9 sec when a spoken word is present. The amplitudes by comparison are lowest during the quiet segment (1.4 sec). Using this approach, we can generate a time-

varying amplitude distribution by repeating these measurements sequentially at consecutive 50 msec intervals (Fig. 3.1 C). To do so, the color on the graph represents the probability of observing particular spectrogram amplitude and the distributions are plotted versus time.

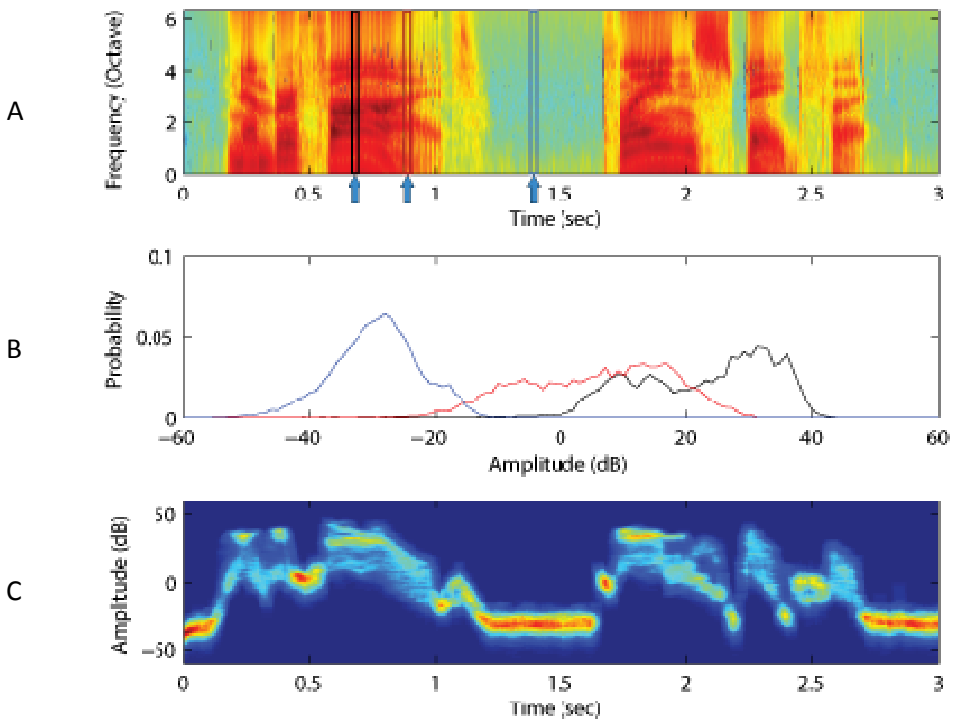

Figure 3.0: A. Spectro-temporal envelope of human speech B. Amplitude distributions at time frames (black=0.7 sec, red= 0.9 sec, blue=1.4 sec) C. Probability distributions

In general, vocalization and speech sounds have non-stationary / time-dependent amplitude distributions. This is seen in the speech example of Fig. 3.0 where the distribution varies between loud (high dB) and soft sound segments (low dB) in a time-dependent manner. Furthermore, note that the width of the amplitude distribution

(measured as a standard deviation, σ_{dB}) also varies with time. This indicates that the instantaneous contrast or equivalently the dynamic range of the auditory spectrogram gradations (within the 50 msec analysis frame) changes in a time-dependent manner. This type of time-dependent behavior is also observed for animal vocalizations (Fig. 3.1). For both, the bird (bald eagle) and primate (pigmy marmoset) the amplitude distributions vary between loud and soft epochs that may have either high or low contrast.

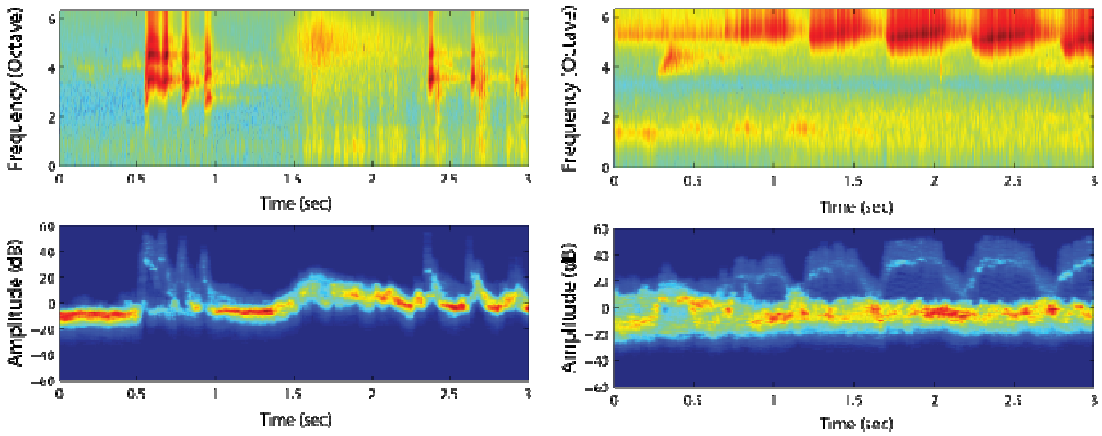

Figure 3.1: Time varying amplitude distributions for animal vocalizations (A) bald eagle (B) pygmy marmoset (new world primate)

By direct comparison, the intensity and contrast statistics of environmental and background sounds can be generally classified as stationary. This is seen for the sounds emanating from a running water source (Fig. 3.2, 3.3 A) and insect chorus (Fig. 3.3 B), both of which have relatively stationary amplitude statistics. That is, the amplitude

distribution is relatively constant at all the time instants and does not fluctuate wildly as for the speech or animal vocalizations. As an example, the amplitude distribution taken at three distinct time-points for the water sound has highly overlapping distributions with highly similar shape (Fig. 3.2). Thus, the amplitude statistics for this water sound are relatively time-invariant and exhibit minimal intensity or contrast fluctuations (i.e., the mean and SD are relatively constant).

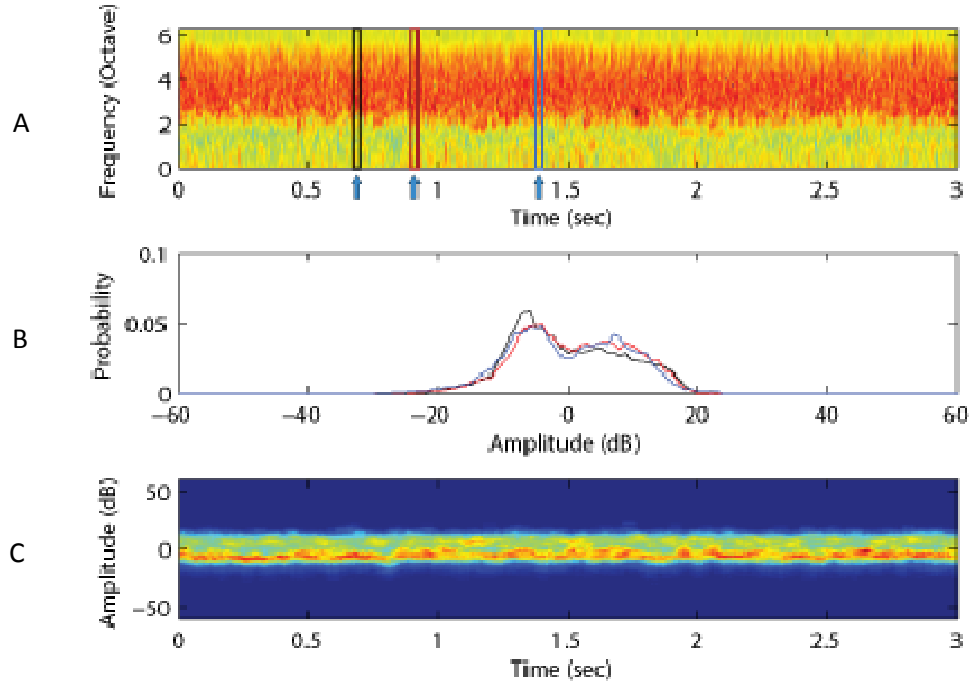

Figure: 3.2. A. Spectro-temporal envelope of a water sound

B. Amplitude distributions at time frames (black=0.7 sec,

red= 0.9 sec, blue=1.4 sec) C. Time-varying amplitude distributions.

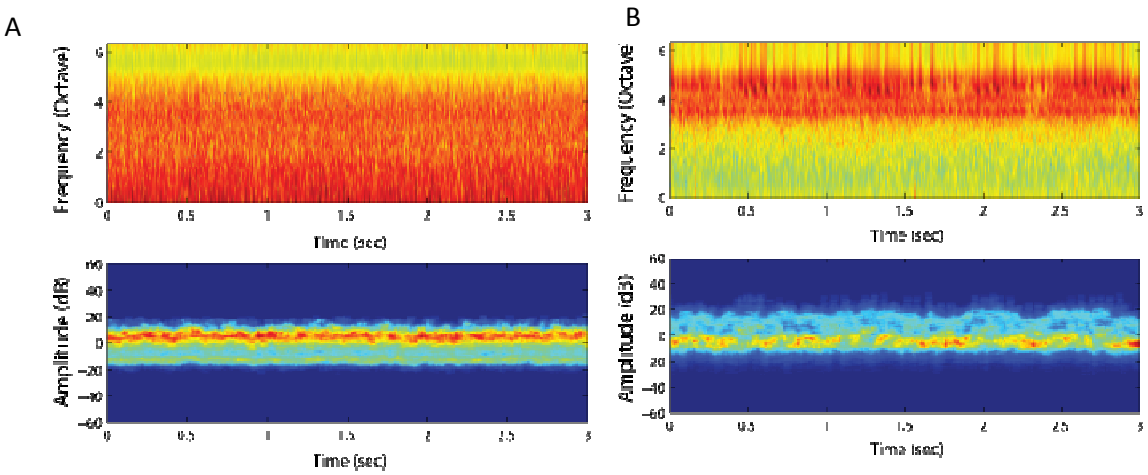

Figure 3.3: Time varying amplitude distributions for background sounds (A) water (B) insects at night

Because the amplitude distribution is a high dimensional description of the spectrogram amplitude fluctuations, we sought to reduce the dimensionality of this statistic. We did so by parameterizing the time-dependent amplitude distribution into a mean (μ_{dB}) and standard deviation (σ_{dB}) value, which we can then plot at each time instant. This is illustrated in Figure 3.4 for a speech sound segment. As can be seen, μ_{dB} and σ_{dB} vary dynamically over time where the mean trajectory follows the center of the amplitude distribution. Fluctuations in the mean of the distribution (μ_{dB}) reflect changes in the mean intensity of the sound whereas fluctuations in the standard deviation (σ_{dB}) reflect changes in the local contrast within a 50 msec sound segment (i.e., the dynamic range). As for speech, animal vocalizations (e.g., parrot, Fig. 3.5 A) exhibit non-stationary statistics such that the instantaneous parameters (μ_{dB} and σ_{dB}) vary dynamically over time. Such behavior was typically not observed for background sounds, where the instantaneous parameters are relative constant over time indicative of stationary contrast and intensity statistics (Fig. 3.5 B).

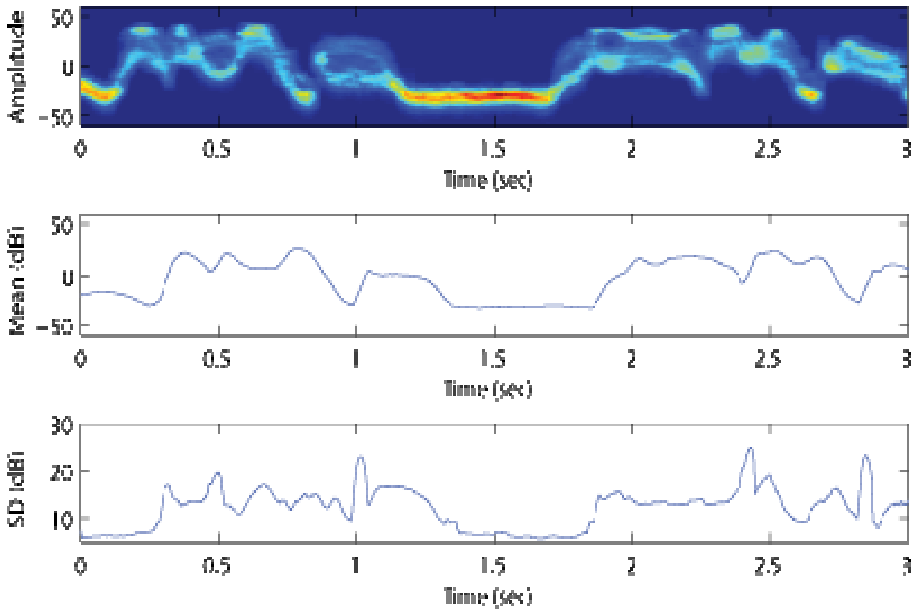

Figure 3.4: Parameterizing amplitude distribution of human speech as time varying parameters, mean and standard deviation.

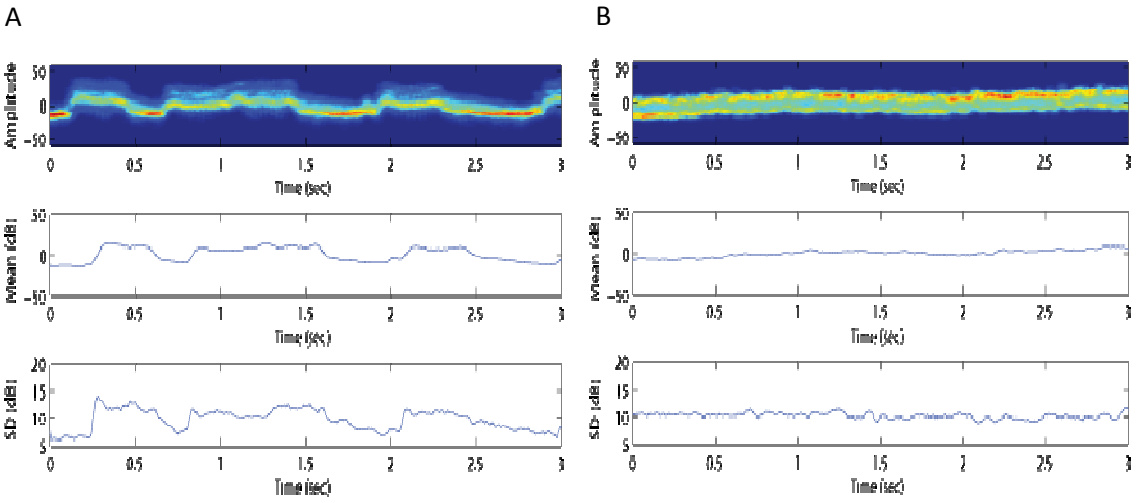

Figure 3.5: Parameterizing amplitude distributions of (A) parrots, (B) water as time varying parameters, mean and standard deviation.

Joint contrast and intensity statistics of natural sound ensembles

The joint intensity and contrast distribution is illustrated for an ensemble of speech (Fig. 3.6) and music (Fig. 3.7). For speech the contrast and intensity span a broad range of values and, in particular, two well isolated and distinct modes can be identified. The first mode occurs for low intensity (low μ_{dB}) and low contrast (low σ_{dB}) whereas a second somewhat more broadly distributed mode is observed for high intensity and contrast. The low intensity-contrast mode occurs during epochs of silence (in between words) and thus corresponds to the background environmental sound. By comparison, the high intensity-contrast mode occurs during periods of speech production. Thus speech has amplitude fluctuations that reflect the contrast statistics of the vocalized speech and the superimposed background sound. Music sounds (Fig. 3.7) and animal vocalizations (Fig. 3.8 A, D) also exhibit broadly distributed intensity-contrast distributions. For instance, primate and bird vocalizations both have relative broad distributions each of which have a unique pattern. For instance, the contrast of primate vocalization extends out to ~20 dB SD while that of birds is somewhat more restricted (mostly <15 dB). However, unlike primate vocalizations intensity and contrast of bird sounds are highly correlated with one another (diagonal orientation). By direct comparison, the range of intensities and contrast for background sounds are substantially more restricted (e.g., Fig. 3.8 B, water) than that of vocalizations. Finally, as a reference, white noise (Fig. 3.8 C) has minimal variability with nearly all of the measurements falling around 6 dB contrast and 0 dB relative intensity.

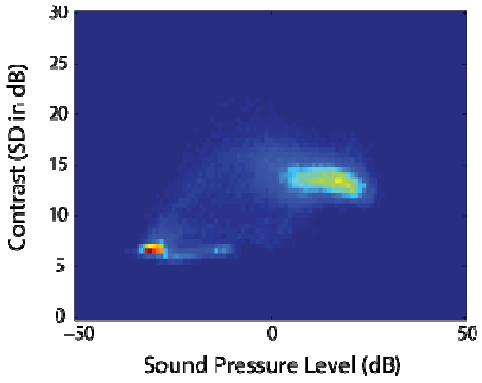

Figure 3.6: Intensity vs Contrast statistics for humans

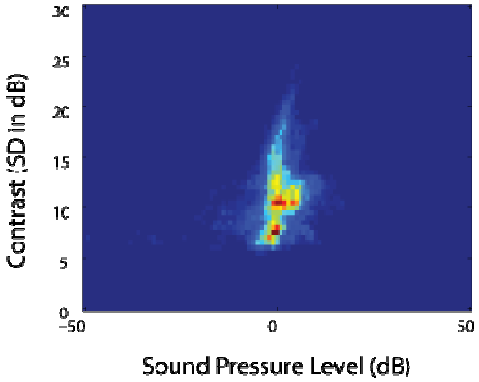

Figure 3.7: Intensity vs Contrast statistics for music (classical)

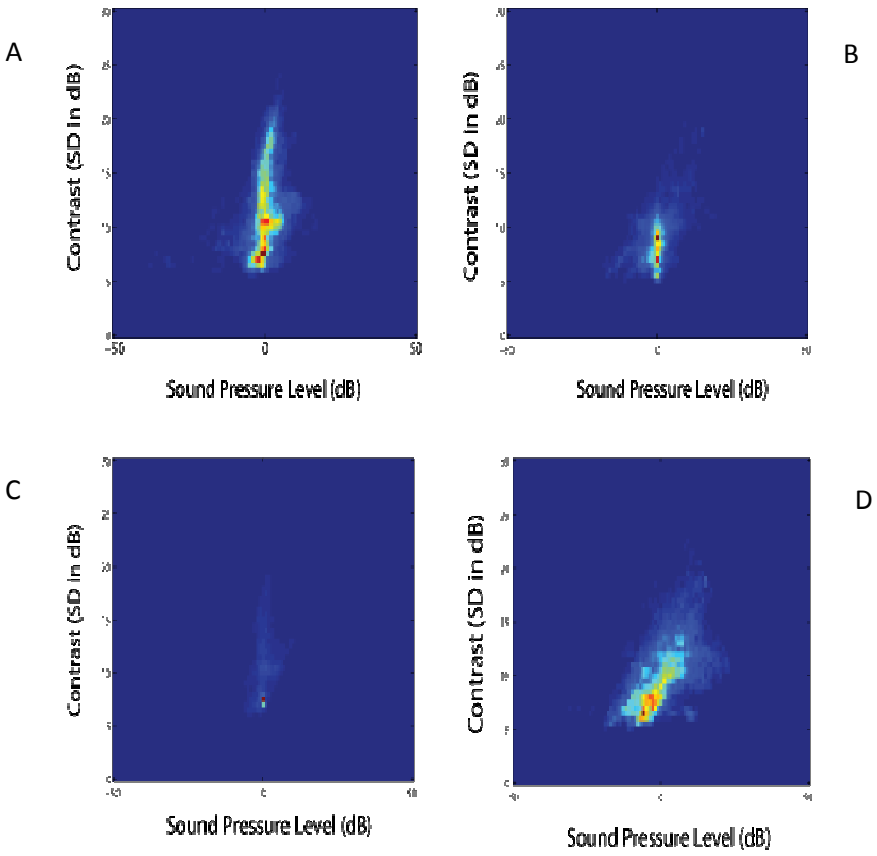

Figure 3.8: Intensity vs Contrast statistics for (A) Primates (B) Water sound (C) White Noise (D) Birds

Discrimination performance depends on observation time

Because the joint intensity and contrast statistics of natural sounds have unique structure that varies from one sound category to another, we hypothesize that the intensity and contrast statistics can be used to categorically discriminate sounds. Using the *prior* distributions defined for each of the sound ensembles above, we used a Naïve Bayesian classifier to measure the discrimination capabilities of the contrast and intensity statistics (see Methods). Discrimination performance was measured by sequentially adding

additional measurements (μ_{dB} and σ_{dB}) across multiple 50 msec time-frames. Below, the classifier performance is shown for parrots vs. cat comparison (Fig. 3.9) and parrots vs. speech (Fig. 3.10). As can be seen, the classifier performance (percent correct classification) is above change (50 %) for both comparisons even when for a single measurement of μ_{dB} and σ_{dB} (i.e., 50 msec observation). The performance of the classifier improves as more observations are included (additional time-frames) reaching near 100% classification rates after measuring 512 time-frames (25.6 seconds of sounds). The performance of Bayes classifier increases with the number of time-frames, reaching perfect value for most sound categories at 256 points (12.8 seconds). This indicates that the joint contrast and intensity statistics have the potential to discriminate amongst sound categories and classification performance improves with observation time.

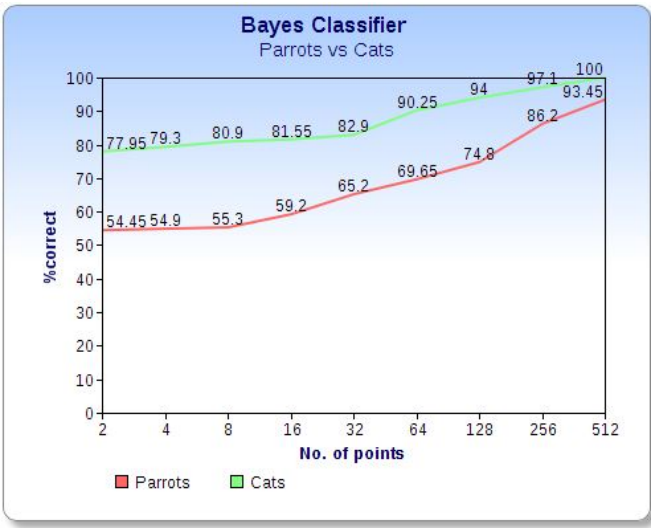

Figure 3.9: Classifier performance increases with number of sample points. Parrots vs Cats

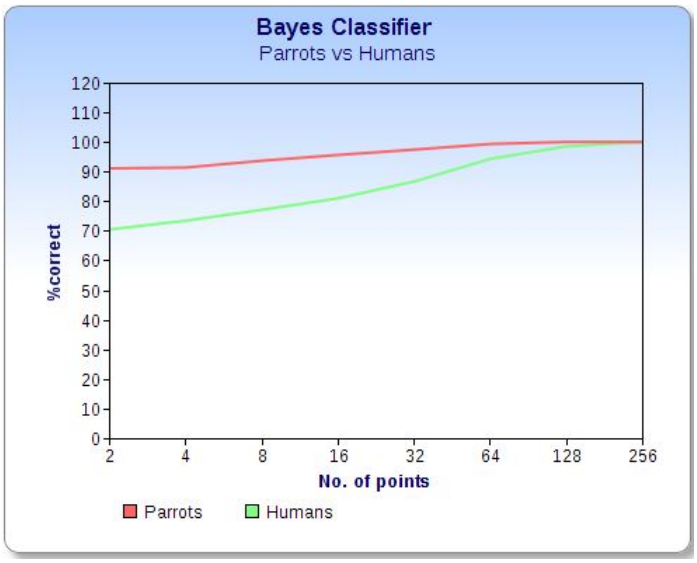

Figure 3.10: Classifier performance increases with number of sample points. Parrots vs Humans

The role of contrast and intensity for sound category discrimination performance

We tested the performance of the Bayesian classifier applied to the contrast and intensity statistics using a 15 alternative forced choice task. A sound from one of the 15 categories was delivered to the classifier and the classifier in turn was required to make a selection of which category the sound originated from. The classifier performance is shown as a confusion matrix (Fig. 3.11 and 3.12) for an experiment in which 32 time-frames (1.6 seconds of sound) where used to categorize sounds. In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instance of the actual sound class that was sent to the classifier, while each row represents the instances of the predicted class by the classifier.

The numerical values indicate the probability of occurrence of occurrence for each of the actual and predicted class combinations. Since we compared 15 categories against each the confusion matrix contains 15 x 15 cells. As can be seen for this example, the highest probability occurs along the diagonal, indicating a correct classification where the actual and predicted class produces a correct match. Since there are a total of 15 sound classes that are delivered at random with equal likelihood to the classifier, the percent of correct identification are well above chance level (1/15 = 6.7%). Thus, despite the relatively high difficulty of this task (15 possible outcomes) the classifier can correctly identify the sound class 60.0% of the times if 1.6 seconds of sound are available.

	Bats	Frogs	Parrots	Humans	New	Water	Old	Squirrels	Birds	Cats	Lions	Dogs	Insects	Music	Tools
Bats	52.9	0	2.2	0	0.7	12.75	0.05	12.55	1.9	0.75	0.3	0.05	14.9	0.15	0.8
Frogs	9.95	43.6	2.05	0.35	0.6	12.45	0.6	6	3.55	4.5	0	7.25	1.45	7.65	0
Parrots	2.4	1.15	39.4	0.8	0	2.2	1.25	5.45	9.75	24.9	1.95	1.7	2.8	3.3	2.95
Humans	1.05	0.45	8.3	76.75	0.2	0	0.95	0.45	0.55	0.75	3.3	6.65	0	0.6	0
New Pri	6.7	4.5	0	0	83.7	0	2.5	0.9	0	0	0.05	0	0	1.65	0
Water	0	0	0	0	0	84	0	5.85	0	0	0	0	4.6	0	5.55
Old Pri	3.3	12.2	0	0	5.15	0	52.75	2.7	0	14.15	0.45	6.15	0	3.15	0
Squirrels	14.65	3	2.35	0.05	0.7	6.3	0.9	64	1.4	2.7	0.6	0.1	1.25	0.5	1.5
Birds	3.75	1.3	10.75	0	0.2	0.8	0.45	3.75	59.4	10.1	2.55	0.3	2.45	3.5	0.7
Cats	2.95	0	1.7	0	2.75	2.6	0	4.4	8.55	65.2	2.9	0	0.05	5.4	3.5
Lions	6.35	4.85	5.6	2.1	3.45	4.75	3.25	3.4	2.1	2.95	36.6	15.2	0.45	7.45	1.5
Dogs	2	1.65	4.15	0.75	5.35	0	10.3	1.15	0.6	3.05	0.7	52.75	0	17.55	0
Insects	7.7	1.9	1.15	0.05	0.1	19.2	0	5.7	0	0.45	1.35	1	56.4	0.75	4.25
Music	0.95	0.2	1.45	0	0	0.55	9	2.3	1.65	35.95	0	0	2.4	39.9	5.65
Tools	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Figure 3.11: Confusion Matrix for 15 sound categories for 32 sample points (1.6 seconds)

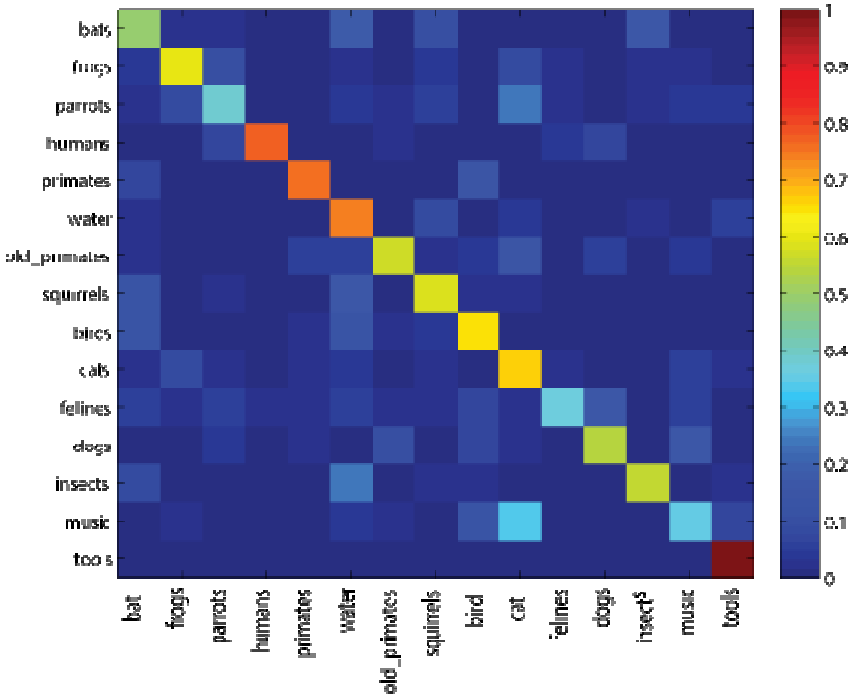

Figure 3.12: Graphical representation of Confusion Matrix for 15 sound categories for 32 sample points (1.6 seconds)

As for the two alternative comparison of Fig. 3.13, the classifier performance improved systematically with increasing sound duration for the 15 alternative forced choice comparisons. The confusion matrix is shown for various sound durations from 100 msec to 12.8 sec (2, 32, 128, 256 time-frames). As can be seen, the confusion matrix becomes increasingly diagonalized such that the percent correct classification increases systematically (40.43%, 60.10%, 79.97%, 87.19% respectively correct classification) with increasing sound duration. Thus, the classifier is capable of reaching nearly perfect classification rates for sound durations in the order of ~10 sec.

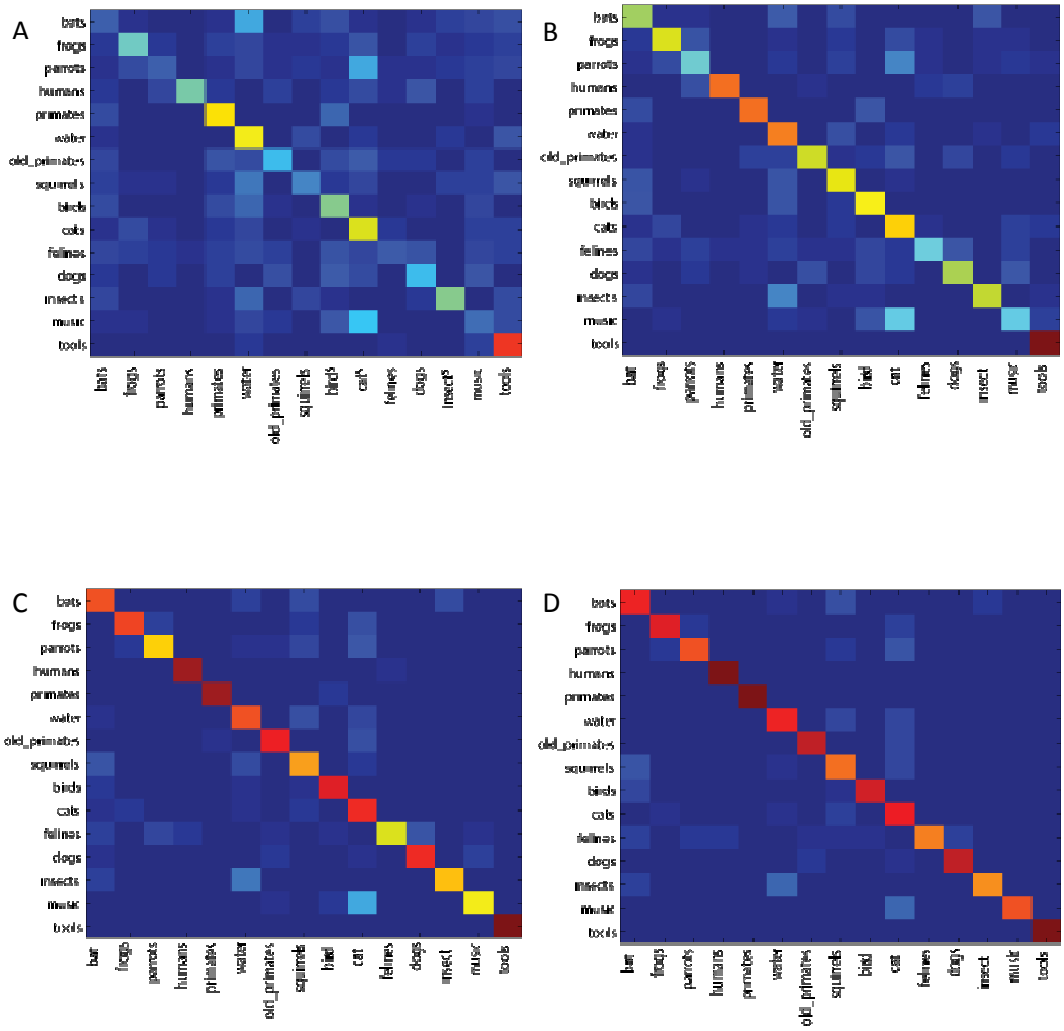

Figure 3.13: Classifier performance increases with number of points (A) 2 (B) 32 (C) 128 (D) 256

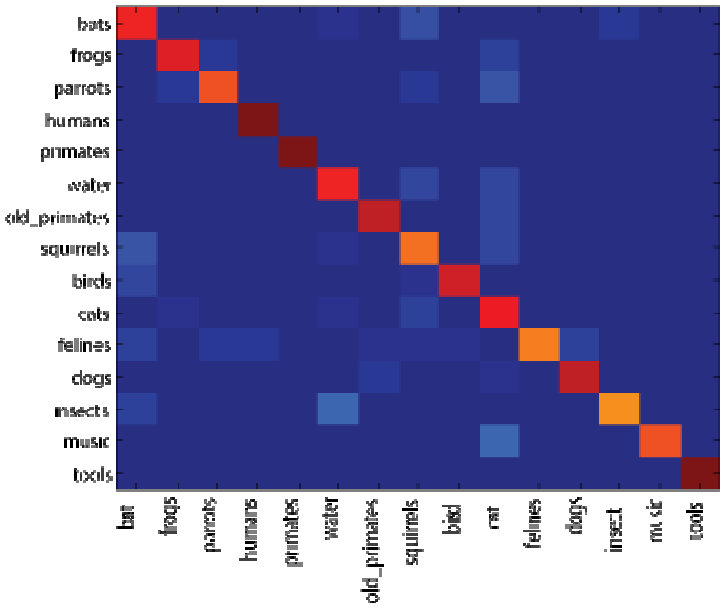

Figure 3.14: Classifier performances at 256 sample points

The results suggest that contrast and intensity statistics are information bearing attributes of natural sound that can aid in sound category identification. Yet, it's unclear how each of these statistics (μ_{dB} or σ_{dB}) individually contribute to sound category discrimination. For this reason, we measured the performance of the Bayesian classifier using individual statistics (μ_{dB} or σ_{dB} alone). Comparing the performance for the joint statistics, the classifier for each statistic performed poorly even for long sound durations (Fig. 3.15, 3.16, shown at 12.8 sec). In fact the performance classifier for mean only and the classifier for standard deviation only with n=256 points was comparable to that of the classifier performance of the joint measurements at n=2 points (100 msec) (Fig. 3.17) This demonstrates that combining statistics leads to substantial increase in the efficiency of sound discrimination and implies that interactions in the joint statistics convey far more information about the sound categories than either statistic alone.

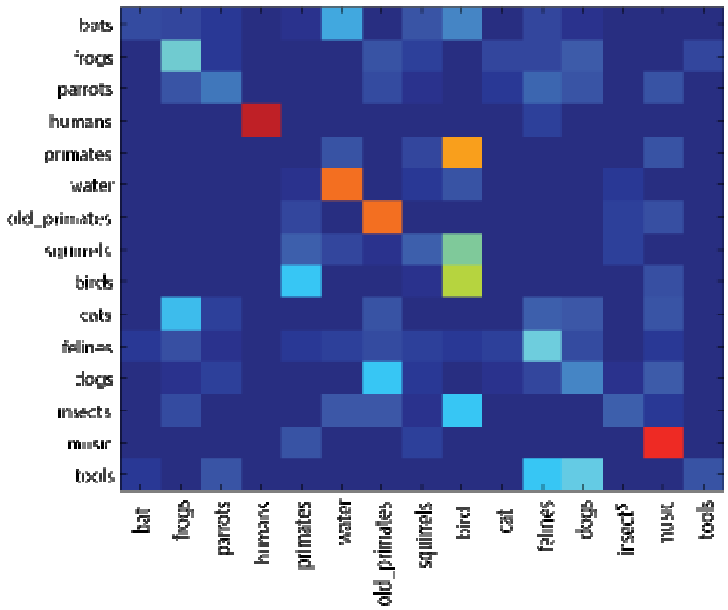

Figure 3.15: Classifier performance for Mean alone at 256 sample points

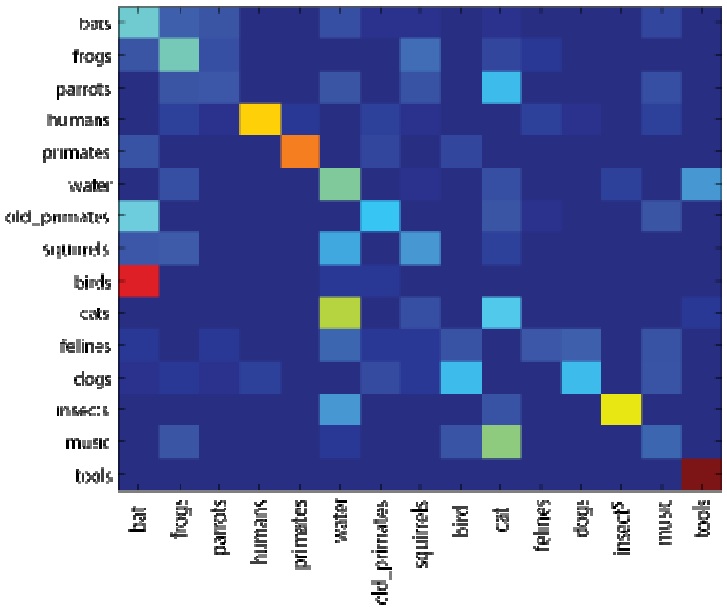

Figure 3.16: Classifier performance for SD alone at 256 sample points

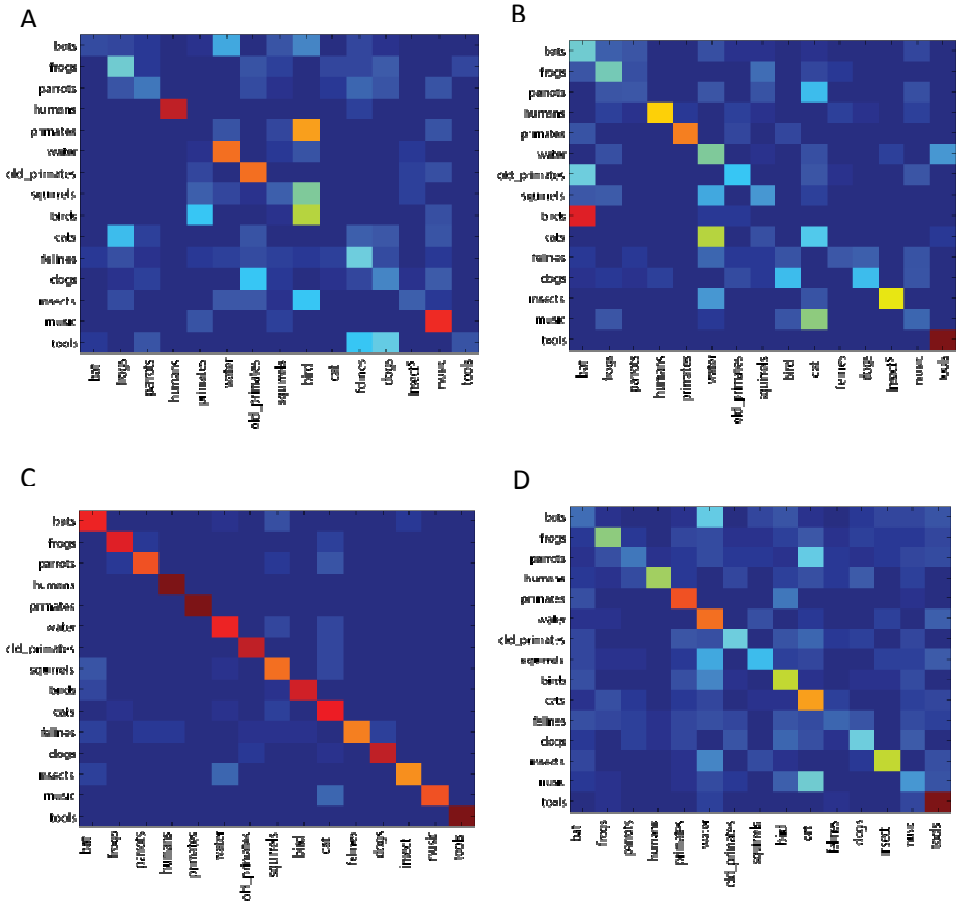

Figure 3.17: Substantial increase in Classifier performance after combining statistics (A) Mean only, 256 points (B) Standard deviation only, 256 points (C) Combined, 256 points (D) Combined, 2 points

CHAPTER 4 – DISCUSSION

We have evaluated the time-varying structure of high-order statistics for a large ensemble of natural sounds and measured their role for sound discrimination and categorization. We tested the hypothesis that biologically motivated high-order sound statistics can enable/enhance discrimination and identification of sound categories from a computational standpoint. The statistical distributions from distinct categories had a unique pattern that enabled discrimination amongst the sound categories tested. Generally speaking, background sounds are stationary and have little variation in their contrast and intensity. By comparison, vocalizations are non-stationary and exhibit substantially more variability.

We used a Naïve Bayes classifier and signal detection theory to identify the role of contrast and intensity statistics. On their own, contrast and intensity contributed to discrimination of sound categories; however, classifier performance was poor for isolated statistics and a substantial improvement in the discrimination performance was observed when these statistical features are measured jointly. The improvement was not simple linear summation of the classifier performance for each statistics as there was a 2.3 fold increase in the correct classification rate when contrast and intensity statistics were jointly included in the classifier. Furthermore, the classification performance was strongly dependent on the observation time interval, such that increasing the observation time leads to improved classification

Prior studies evaluated the role of high-order sound statistics in human observers using stationary texture sounds [4, 14]. These studies demonstrated that stationary high-order statistics can contribute to identification and discrimination for sounds with stationary statistics. Yet, many man-made and natural sounds, such as animal vocalizations or music are non-stationary and texture synthesis models fail to replicate their sound properties. Our results add to these findings since they suggest that time-varying statistics of contrast and intensity contain substantial information that enables discrimination amongst sound categories. Thus it is feasible that such non-statistics could be incorporated into predictive sound synthesis and compression algorithms. Furthermore, although this study examined the role of time-varying statistics from a strictly computational standpoint it demonstrates that there is substantial time-varying information that humans and animal can potentially exploit for sound recognition, discrimination, and source segregation tasks.

This study explicitly tested the role of time-varying contrast and intensity statistics for discrimination of sound categories. It is likely that non-stationary structure for other high-order sound statistics can contribute to sound recognition and discrimination phenomena. For instance, across-channel correlations in sounds are non-stationary [6, 15] and can potentially improve signal detection in noise. In general, across-frequency correlations are non-stationary for vocalization and sounds and thus it is likely that such time-varying statistics could enhance signal detection. Non-stationary statistics in the frequency correlation structure of vocalizations could theoretically aid in the detection of signals within the presence of stationary background noises. Thus the role of

other high-order statistics and the corresponding time-varying structure needs to be critically evaluated in future studies.

Although it is clear that the auditory system utilizes such statistics from a perceptual standpoint, it is unclear how such statistics are computed or extracted from real world sounds by the brain. Neurons in the auditory midbrain and cortex can respond selectively to contrast and intensity statistics[1, 5, 7, 8, 16], providing plausible mechanisms for how such features might be extracted by the brain. Since central auditory neurons rapidly adapt to statistics of natural stimuli[7-10, 16], it is also plausible that adaption provides a mechanism for the brain to efficiently track sound statistics over perceptually relevant time-scales. Thus, future studies need explicitly to test the hypothesis that brain computes and extracts such sound statistics for sound recognition and discrimination tasks. Ultimately, a comprehensive theory for understanding the role sounds statistics play needs to consider the acoustic structure of real world sounds (natural and man-made), the role of neural computational mechanisms, and ultimately their relationship to behavioral performance.

REFERENCES

1. Escabi, M.A., Miller, L.M., Read, H.L. & Schreiner, C.E. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J Neurosci* **23**, 11489-11504 (2003).
2. Rodriguez, F.A., Chen, C., Read, H.L. & Escabi, M.A. Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *J Neurosci* **30**, 15969-15980 (2010).
3. Singh, N.C. & Theunissen, F.E. Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* **114**, 3394-3411 (2003).
4. McDermott, J.H. & Simoncelli, E.P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* **71**, 926-940 (2011).
5. Barbour, D.L. & Wang, X. Contrast tuning in auditory cortex. *Science* **299**, 1073-1075 (2003).
6. Nelken, I., Rotman, Y. & Bar Yosef, O. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **397**, 154-157 (1999).
7. Rabinowitz, N.C., Willmore, B.D., Schnupp, J.W. & King, A.J. Contrast gain control in auditory cortex. *Neuron* **70**, 1178-1191 (2011).
8. Rabinowitz, N.C., Willmore, B.D., Schnupp, J.W. & King, A.J. Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J Neurosci* **32**, 11271-11284 (2012).
9. Kvale, M.N. & Schreiner, C.E. Short-term adaptation of auditory receptive fields to dynamic stimuli. *J Neurophysiol* **91**, 604-612 (2004).
10. Nagel, K.I. & Doupe, A.J. Temporal processing and adaptation in the songbird auditory forebrain. *Neuron* **51**, 845-859 (2006).
11. Aide, T.M., C. Corrada-Bravo, M. Campos-Cerqueira,C. Milan, G. Vega and R. Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ* **1:e103; DOI 10.7717/peerj.103** (2013).
12. Shapley, R.M. & Victor, J.D. The effect of contrast on the transfer properties of cat retinal ganglion cells. *J Physiol* **285**, 275-298 (1978).
13. Shapley, R. & Reid, R.C. Contrast and assimilation in the perception of brightness. *Proc Natl Acad Sci U S A* **82**, 5983-5986 (1985).
14. McDermott, J.H., Schemitsch, M. & Simoncelli, E.P. Summary statistics in auditory perception. *Nat Neurosci* **16**, 493-498 (2013).
15. Escabi, M.A. & Schreiner, C.E. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *J Neurosci* **22**, 4114-4131 (2002).
16. Dean, I., Robinson, B.L., Harper, N.S. & McAlpine, D. Rapid neural adaptation to sound level statistics. *J Neurosci* **28**, 6430-6438 (2008).